Measures of Central Tendency: Mean, Median, and Mode

Central tendency is a statistical measure that defines the center of a distribution. The goal is to identify a single value that best represents an entire set of data. The three most common measures are the Mean, Median, and Mode.

1. The Mean (The Average)

The Mean is the most widely used measure of central tendency. It is calculated by summing all the values in a dataset and dividing the sum by the total number of values.

Definition and Calculation

- **Definition:** The arithmetic average of all values.
- **Calculation:** $\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$

When to Use the Mean

The Mean is the best choice when:

Scenario	Explanation
Symmetrical Data	The data is normally or symmetrically distributed (bell-curve shape).
Interval/Ratio Data	The data is measured on an interval or ratio scale (e.g., temperature, height, salary).
Further Analysis	You plan to conduct more advanced statistical tests (e.g., standard deviation, t-tests, linear regression).
Equal Importance	You assume all values in the dataset are equally important and representative.

Caution: The Mean is heavily influenced by **outliers** (extreme values). In a highly skewed dataset (like income), the Mean can be misleadingly high or low.

2. The Median (The Middle)

The Median is the middle value in a dataset when the values are arranged in ascending or descending order.

Definition and Calculation

- **Definition:** The value that separates the higher half of the data from the lower half.
- **Calculation:** Sort the data. If n is odd, the median is the center value. If n is even, it is the average of the two center values.

When to Use the Median

The Median is the best choice when:

Scenario	Explanation
Skewed Data/Outliers	The dataset is highly skewed, or there are significant outliers (e.g., real estate prices, highly variable wait times).
Ordinal Data	The data is measured on an ordinal scale (e.g., survey responses like "very satisfied," "satisfied," "neutral").
Unequal Spacing	The distribution is open-ended or has unequal intervals, making the mean unreliable.

Benefit: The Median is robust—it is **resistant to outliers** and provides a truer representation of the "typical" value in skewed distributions.

3. The Mode (The Most Frequent)

The Mode is the value that occurs most frequently in a dataset. A dataset can have one mode (unimodal), two modes (bimodal), or more than two modes (multimodal).

Definition and Calculation

- **Definition:** The most frequently occurring value or category in the dataset.

- **Calculation:** Simply count the frequency of each distinct value or category.

When to Use the Mode

The Mode is the best choice when:

Scenario	Explanation
Nominal/Categorical Data	The data is non-numeric or purely categorical (e.g., favorite color, type of car, ethnicity, gender).
Discrete Data	You need to identify the most popular or common choice among a set of discrete options (e.g., the most sold shoe size, the highest vote-getter).

Benefit: The Mode is the **only measure** that can be used for purely categorical (nominal) data.

Summary of Use Cases

Measure	Best for Data Type	Best for Distribution	Sensitivity to Outliers
Mean	Interval, Ratio	Symmetrical/Normal	High (Easily affected)
Median	Ordinal, Interval, Ratio	Skewed (Non-Normal)	Low (Resistant)
Mode	Nominal, Discrete	Any, especially categorical	None